Published online February 2, 2006

Use of the Nonparametric Nearest Neighbor Approach to Estimate Soil Hydraulic Properties

Attila Nemes,* Walter J. Rawls, and Yakov A. Pachepsky

ABSTRACT

Nonparametric approaches are being used in various fields to address classification type problems, as well as to estimate continuous variables. One type of the nonparametric lazy learning algorithms, a k-nearest neighbor (k-NN) algorithm has been applied to estimate water retention at −33- and −1500-kPa matric potentials. Performance of the algorithm has subsequently been tested against estimations made by a neural network (NNet) model, developed using the same data and input soil attributes. We used a hierarchical set of inputs using soil texture, bulk density (D_b), and organic matter (OM) content to avoid possible bias toward one set of inputs, and varied the size of the data set used to develop the NNet models and to run the k-NN estimation algorithms. Different 'design-parameter' settings, analogous to model parameters have been optimized. The k-NN technique showed little sensitivity to potential suboptimal settings in terms of how many nearest soils were selected and how those were weighed while formulating the output of the algorithm, as long as extremes were avoided. The optimal settings were, however, dependent on the size of the development/reference data set. The nonparametric k-NN technique performed mostly equally well with the NNet models, in terms of root-mean-squared residuals (RMSRs) and mean residuals (MRs). Gradual reduction of the data set size from 1600 to 100 resulted in only a slight loss of accuracy for both the k-NN and NNet approaches. The k-NN technique is a competitive alternative to other techniques to develop pedotransfer functions (PTFs), especially since redevelopment of PTFs is not necessarily needed as new data become available.

MODELING WATER and solute transport has become an important tool in simulating agricultural productivity as well as environmental quality. The use of models, however, is often limited by the lack of information on soil hydraulic properties. For many applications, the estimation of those properties using PTFs is a feasible alternative to the costly and time-consuming measurements.

Regression techniques and lately artificial NNets are two of the most commonly used methods to develop PTFs. One common feature of today's PTFs is that they are all based on the parametric approach, that is, they are equations with parameters found from fitting those equations to data, which has several drawbacks. Identifying the right equation, and ensuring that the associated probability distributions of errors will be similar across the data space is not always easy. Estimation results can be heavily biased in case of small sample sizes. The equations need to be redeveloped and republished, should new data become available, and users are not able to simply include any additional data sets to improve performance for their site-specific range of soil properties.

An alternative approach for such estimations could be the use of nonparametric techniques. Such techniques are based on pattern-recognition and using similarities rather than on fitting equations to data. Use of a nonparametric algorithm is beneficial when the form of relationship between input and output data is not known a-priori (Yakowitz, 1993; Lall and Sharma, 1996). Such is the case with soil hydraulic properties, where the form of their dependence on other soil attributes is not known in advance.

Nonparametric methods are being applied in several fields of hydrology. One of such approaches is the k-NN approach. The k-NN classification techniques have been developed and applied in many papers in the fields of pattern recognition and statistical classification (Dasarathy, 1991). The approach can be found in the literature of many fields besides hydrology, for example, forestry, plant physiology, virology, molecular biology, entomology, zoology, agronomy, and biochemistry. Examples for applications relevant to hydrologic simulation include stream flow simulation and disaggregation using nearest neighbor and kernel methods (Lall and Sharma, 1996; Sharma et al., 1997; Tarboton et al., 1998; Kumar et al., 2000; Sharma and O'Neill, 2002, Souza Filho and Lall, 2003), simulation of rainfall using a nonhomogeneous Markov chain model (Rajagopalan et al., 1996; Sharma and Lall, 1999; Marshall et al., 2004), simulation of rainfall spells using a seasonally homogeneous resampling technique (Lall et al., 1996), flood forecasting (Sankarasubramanian and Lall, 2003), and the simulation of multivariate daily weather sequences (Rajagopalan et al., 1997; Yates et al., 2003; Harrold et al. 2003a, 2003b; Clark et al., 2004). Yakowitz and Karlsson developed a theoretical basis for using k-NN methods for time series forecasting and applied them in a hydrologic context (Karlsson and Yakowitz, 1987; Yakowitz and Karlsson, 1987; Yakowitz, 1993).

Such approaches have not yet been widely used in studies related to unsaturated soil hydrology. A similarity-based k-NN type technique has been applied successfully

A. Nemes, Dep. of Environmental Sciences, Univ. of California, Riverside, CA 92521; A. Nemes and W.J. Rawls, USDA-ARS Hydrology and Remote Sensing Lab., 10300 Baltimore Ave., Bldg. 007, BARC-West, Beltsville, MD 20705; Y.A. Pachepsky, USDA-ARS Environmental Microbial Safety Lab., Powder Mill Road, Bldg. 173, BARC-East, Beltsville, MD 20705. Received 21 Apr. 2005. *Corresponding author (anemes@hydrolab.arsusda.gov).

Published in Soil Sci. Soc. Am. J. 70:327–336 (2006).
Soil Physics
doi:10.2136/sssaj2005.0128
© Soil Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: D_b, bulk density; k-NN, k-nearest neighbor technique; MR, mean residual; NNet, neural network; OM, organic matter; PTF, pedotransfer function; RMSR, root-mean-squared residual; SSC, sand, silt, and clay contents (soil texture); WRC, water retention curve.

to interpolate soil particle-size distributions by Nemes et al. (1999). They found this technique to perform well while estimating the missing 50-μm particle fraction for many European soils, which later served as input to soil hydraulic PTFs. Jagtap et al. (2004) introduced a dynamic k-NN approach to estimate the drained upper limit and lower limit of plant water availability from field measured soil water retention information. They compared their model to existing soil hydraulic PTFs to make estimations for their data set.

The k-NN technique and many of its derivatives belong to the group of 'lazy learning algorithms'. It is 'lazy', as it passively stores the development data set until the time of application; all calculations are performed only when estimations need to be generated. Application of this technique means identifying and retrieving the nearest (most similar) stored objects to the target object. The quality of such estimations depends on, among others, which objects are ruled to be the nearest to the target object. One concern is that a standard k-NN does not perform attribute selection; it allows irrelevant or interacting inputs to have as much effect on the distance calculation as any other useful inputs. Another concern is that some inputs may have a (considerably) wider range of data than others. A unit change in one input variable may have much larger influence on the distance measure than the same change in the other. Such concerns lead to the introduction of data normalization systems and different attribute weighing systems in more recent k-NN variants (e.g., Wettschereck et al., 1997).

There are many studies in literature that evaluate and compare the performance of soil hydraulic PTFs. A limitation of many of such studies is that it remains unclear what the main sources of the differences in estimation errors are. It is not clear whether differences between data sets used to derive PTFs (size, origin, reliability), differences between the algorithms of PTF development (e.g., different regression types vs. NNet models) or differences among the used input attributes cause a particular PTF to perform better than others. Many PTF comparison studies have no particular goal to promote any particular method or data source over others (e.g., Tietje and Tapkenhinrichs, 1993; Kern, 1995; Imam et al., 1999; Cornelis et al., 2001; Gijsman et al., 2003), they simply compare the performance of PTFs available in literature on some specific data. Other studies promote the advantages of using a particular data set or method over others (e.g., Pachepsky et al., 1996; Nemes et al., 2003; Jagtap et al., 2004). When a novel approach to make estimations is introduced, all other factors (e.g., differences in development data and/or inputs used) should be eliminated to allow reporting any advantage gained solely due to using the novel approach. The study of Jagtap et al. (2004) compares the performance of k-NN estimations to other PTFs that were developed using different methods, using different data, and using different input attributes at the same time. The performance of PTFs varies with the pedological origin of the soils on which they were developed. The effect of using different data sets on the soil water retention estimations has been pointed out by different authors (e.g., Rawls et al., 1991; Schaap and Leij, 1998; Minasny et al., 1999; Nemes et al., 2003), and the effect of using different inputs in PTFs on estimations has been shown by for example, Schaap et al. (2001) and Nemes et al. (2003).

The objective of this study was to introduce the k-NN approach for the estimation of water retention points and to compare its performance to the performance of NNet models in terms of estimation accuracy. For the reasons listed in the previous paragraph, we developed NNet PTFs using the same data sets and the same inputs as with the k-NN approach. In this study, we introduce a relatively simple form of the k-NN approach to examine the worthiness of this technique for the estimation of soil hydraulic data.

MATERIALS AND METHODS

Soil Data

There were 2125 soil horizons selected from the U.S. NRCS-SCS Soil Characterization Database (Soil Survey Staff, 1997), according to the following criteria: Mineral soil horizons were selected from the contiguous USA having horizon notation 'A', 'A1', and 'Ap' (and their derivatives), with the condition that the top of the horizon was at the soil surface. Organic matter (OM) content of the selected soils was limited to 1 to 15%, and D_b was limited to 0.5 to 2.0 g cm^{-3}. Selected soil properties were the following: sand (50–2000 μm), silt (2–50 μm), and clay content (<2 μm) according to the USDA classification system (USDA, 1951), D_b, OM content and retained (volumetric) water at −33- and −1500-kPa matric potentials (θ33 and θ1500 respectively), with no missing data allowed in any of the fields. Such matric potentials were chosen as those are often used to approximate field capacity and wilting point when calculating plant available water, and thus are often preferred points in water retention curve (WRC) determinations in the laboratory. Measured WRC data at those matric potentials can be found frequently in many soil hydraulic databases worldwide. The 2125 size data set excludes any entries that showed obvious inconsistency in physical and/or hydraulic data (sand + silt + clay $\neq$ 1; $\theta 33 < \theta 1500$; $\{[1 - (D_b)/2.65] - \theta\ 33\} < 0$).

Table 1 shows the summary statistics of selected soil attributes of the selected data set. The data set contains data of a wide range of soils, in terms of the shown soil attributes. We note the unusually high maximum value for −1500-kPa water content. Water contents in the NRCS database are stored as gravimetric water content. Different D_b values are stored in the database- measured at different state of wetness that have to be used to convert −33 and −1500 kPa gravimetric water contents to their respective volumetric water content values. This resulted in a few large −1500 kPa volumetric water content values in case of some—presumably

Table 1. Summary statistics of selected soil attributes in the source data set used to develop pedotransfer functions.

Properties	Unit	MIN	MAX	AVG	SD	MEDIAN
USDA sand	g g^{-1}	0.004	0.955	0.280	0.231	0.211
USDA silt	g g^{-1}	0.034	0.922	0.492	0.194	0.491
USDA clay	g g^{-1}	0.002	0.811	0.228	0.133	0.205
Bulk density	g cm^{-3}	0.520	1.890	1.362	0.186	1.380
Organic matter	%	1.000	14.861	3.082	2.063	2.500
θ (−33 kPa)	m m^{-3}	0.051	0.724	0.316	0.083	0.325
θ (−1500 kPa)	m m^{-3}	0.022	0.714	0.171	0.094	0.153

swelling—soils with high clay content. We converted gravimetric water contents to volumetric water contents to remain compatible with most existing PTFs.

The data set has been used as development data as well as to provide data to test the estimations. Samples have been randomly drawn to be either member of the development data set (reference data set for the k-NN technique) or of the test data set. We elected to use 435 samples as test data. We used different sizes of development/reference data sets to evaluate the effect of the size of the development data set on the two methods that are compared. Samples were drawn to be members of development/reference data sets of 100, 200, 400, 800, and 1600 samples. All random data selections have been repeated 200 times to allow the development of an ensemble of PTF estimations. By using a sufficiently large number of replicates one can minimize the impact of any single replicate (i.e., any particular data set division) on the final estimation results. It has been pretested that, for this application, using 200 replicates is sufficient to make the effect of any single replicate on the estimations insignificant. That number, however, has not been minimized/optimized. Statistical measures in this study to evaluate and compare the two methods are thus based on 200 replicates using each method. Development of such ensemble of estimations has a notable advantage over single PTF estimations, that is, the uncertainty of estimates can be quantified, which can then be subject to statistical analyses and/or be inputted into simulation models.

Rawls et al. (1991) and Wösten et al. (2001) lists the input attributes used by many PTFs. We have chosen to use inputs that are most often used by different authors: sand, silt, and clay content (SSC), D_b, and OM content. We assumed that they are all equally relevant and important in the estimation of the output attributes. Four different sets of input attributes were used to estimate water retention at two different matric potentials (-33 and -1500 kPa) from data of the NRCS data set. The simplest model used only SSC as predictors. In the following two models, either D_b, or OM content was added to SSC as a predictor. In the fourth model all of these inputs were used as predictors. This is to avoid a possible bias while applying one particular set of input attributes, and to account for the situation of different levels of data availability for potential future use or comparisons.

The k-Nearest Neighbor Technique

Unlike classic PTFs, the k-NN technique does not use any predefined mathematical functions to estimate a certain attribute. A ‘reference’ data set—analogous to the development or training data sets used to develop classic PTFs—is searched for soils that are most similar to the target soil, based on the selected input attributes. Apparently, performance of such technique largely depends on the goodness of selection of the ‘most similar’ (nearest) soils. In most k-NN studies, the ‘distance’ measure is calculated as the classical Euclidean distance between the target and the known instances. In a simple case, with only two input attributes, for example, sand and clay content, selection of the nearest (or most similar) soil(s) can be represented geometrically using Pythagoras’ theorem, as demonstrated by Jagtap et al. (2004). The ‘distance’, of each soil from the target soil can be calculated as the square root of the sum of squared differences in sand and clay content between the target soil and each of the soils of the reference data set. Soils of the reference data set will then be sorted in ascending order of their distance from the target soil. The estimated value of the output attribute is calculated as the weighed average of the output attribute of a preselected number of the nearest soils.

Of course, the above case is largely simplified in several aspects. One of the factors that need attention is the fact, that most PTFs utilize information of more than two input attributes. For such cases the generalized form of

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}^2} \qquad [1]$$

may provide sufficient solution, where d_i is the ‘distance’ of the ith soil from the target soil, and Δa_{ij} represents the difference of the ith soil from the target soil in the jth soil attribute. The term ‘distance’, does not refer to actual (physical) distance, but to a measure of similarity; the distance will be smaller for soils that are more similar to the target soil in their input attributes.

A rightly concern is that a unit difference in one attribute may not be as significant as the same unit difference in another attribute, because of differences in the order of magnitude and/or range of data of the different input attributes. For example, sand content, if given in a percentage, can take up values anywhere between 0 and 100, whereas OM content ranges from 0 to a maximum of 15% in nonorganic soils. A unit difference in OM is expected to be more significant than the same unit difference in sand content. To avoid bias toward one attribute or the other, the data need to be normalized before it is used to calculate ‘distance’ using Eq. [1]. In this study we first transformed all input attributes to obtain temporary variables with distribution having zero mean and standard deviation of 1 using the following transformation:

$$a_{ij(temp)} = [(a_{ij}) - \overline{a}_j]/\sigma(a_j) \qquad [2]$$

where a_{ij} represents the value of the jth attribute of the ith soil, $\overline{a}_j$ and $\sigma(a_j)$ represent the mean and standard deviation of the observed values of the jth attribute in the reference data set. Then, we examined the minimum-maximum range of those temporary variables, and scaled the temporary variables to obtain zero mean and the same minimum-maximum range in the data of all attributes:

$$a_{ij(trans)} = a_{ij(temp)}\{MAX[range(a_{j=1(temp)}),\ldots,\\ range(a_{j=x(temp)})]\}/range(a_{j(temp)}) \qquad [3]$$

where $a_{j(\mathrm{temp})}$ represents the data of the jth soil attributes normalized using Eq. [2]; and $\mathrm{a}_{ij(\mathrm{trans})}$ represents the final transformed value of the jth attribute of the ith soil that are to be used as input. Certainly, this method is somewhat sensitive to the potential presence of outliers in any of the attributes, as that may stretch the min-max range of the particular attribute, values of which may then be somewhat under-weighed in Eq. [3], compared with the other attributes.

An additional issue that needs to be addressed is the number of soils (k) to be selected from the reference data set that are then used to formulate the estimate of the output attribute of the target soil. It is not straightforward to tell, whether the closest single soil, the closest two, three, ten, twenty, or perhaps more will give the most accurate estimate. To determine the optimal value of k, the leave-one-out cross-validation technique was used by Lall and Sharma (1996). They suggested a potential choice of $k = n^{1/2}$ for $n > 100$, based on their experience under certain conditions, where n is the number of known instances in the reference data set. They also note that sensitivity of the technique in terms of an accuracy measure, the generalized cross validation score, to the choice of k is small. As we had no prior information about the optimal k for the presented type of application, optimization of k is part of the presented research.

Finally, one has to decide how to weigh each selected soil while forming the estimate of the output attribute, if the selected number of soils is more than one. As a solution, the simple average of their output attribute can be calculated. However, the calculated 'distance' of each soil from the target (see Eq. [1]) will be different, and it can be argued that a soil closer to the target should have more weight in calculating the estimated value for the target. A weighing system that allows distance-dependent weighing of soils seems to offer an alternative solution. One of the possible solutions mentioned in literature is to establish some type of inverse relationship between such weights and the distance of the target from the selected nearest neighbors. For example, Lall and Sharma (1996) applied a system in which weights for each selected neighbor are calculated as:

$$w_{(i)} = \frac{1/i}{\sum_{l=1}^{k} 1/l} \quad [4]$$

where $w_{(i)}$ is the weight associated with the ith nearest neighbor and k is the number of neighbors considered. This approach, however, considers only the rank of each sample in being the nearest neighbor to the target, and does not consider the relative distances of the selected k neighbors from the target. We used a weighing system that accounts for the distribution of the distances of the selected nearest neighbors from the target. Weights for each selected neighbor are calculated as:

$$w_i = d_{i(rel)} / \sum_{i=1}^{k} d_{i(rel)} \quad [5]$$

where k is the number of neighbors considered; w_i is the assigned weight, and $d_{i(rel)}$ is the relative distance of the ith selected neighbor, calculated as

$$d_{i(rel)} = \left(\sum_{i=1}^{k} d_i / d_i \right)^p \quad [6]$$

where k is the number of neighbors considered; d_i is the distance of the ith selected neighbor calculated using Eq. [1], and p is a power term that is to be optimized as part of the present study. The p term is introduced to account for different possible weight/distance relationships. If $p = 1$, a simple inverse relationship is assumed, $p = 2$ assumes inverse squared relationship, etc. We examine what power term (p) could be best used to convert distance to weight.

The Artificial Neural Network Technique

Recently, artificial NNet models have been used successfully in PTF development (e.g., Pachepsky et al., 1996; Tamari et al., 1996; Schaap et al., 1998; Koekkoek and Booltink, 1999; Minasny et al., 1999; Schaap and Leij, 2000; Minasny and McBratney, 2002; Nemes et al., 2003). Most studies found that the predictive capabilities of NNet PTFs were equivalent or superior to different regression type PTFs. For this reason, we have chosen the NNet technique to serve as the basis for comparison for the newly introduced k-NN technique.

A NNet model consists of many simple computing elements (termed neurons or nodes) that are organized into subgroups (layers) and are interconnected as a network by weights. A model typically consists of an input layer, an output layer, and one (or more) 'hidden' layer(s) that connect(s) the input and output layers. The number of nodes in the input and output layers correspond to the number of input and output variables of the model, the number of hidden nodes can be varied freely. Data flow goes from the input layer through the hidden layer(s) to the output layer. A node in the hidden and output layers receives multiple inputs—typically from all nodes of the previous layer. Within the node, each input is weighted and combined to produce a single value as the output of that node, which is then directed to all the nodes of the next layer, or outputted if it was a node of the output layer. The weight matrices are obtained through a calibration (training) procedure, which can then be used to make estimations on independent data. For a more thorough description on NNets, we refer the reader to Hecht-Nielsen (1990) or Haykin (1994).

Following Nemes et al. (2003), we used a three-layer backpropagation NNet model. As the problem to be solved is relatively simple we used one hidden layer only. There are significantly different approaches to set the number of nodes in the hidden layer. We set it empirically as half of the sum of input and output variables, rounded up to the nearest integer. While NNets are able to extract essential information from raw input data, the resulting networks may be complex and the required computation times very long. To reduce both, we transformed all data, before being presented to the NNets, to take up the interval [0,1].

The NNets were combined with the data selection procedure of the bootstrap method (Efron and Tibshirani, 1993) to generate internal calibration-validation data set pairs for an early stopping procedure. The bootstrap method is a nonparametric technique that simulates alternative (replica) data sets out of a single data set. Given a data set of size N, the bootstrap method generates replica data sets, also of size N, by random selection with replacement. Some samples are included more than once, while others are not selected into a particular replica data set. The replica data set is used to calibrate the NNet model while data not in the replica data set are used for validation to stop the calibration process when a minimum error is reached. Multiple realizations of subsets can help to avoid bias toward any particular calibration-validation data set pairs. We generated 10 replica data sets, each of which was used to calibrate the NNet models, which procedure provided 10 'subestimates', that could be slightly different from each other. The estimate of a PTF from one particular data set—for each value—was then calculated by averaging the 10 subestimates of the value. Application of the bootstrap technique took place internally in the NNet program to derive the best estimates from each development data set, and was performed independently within each of the 200 PTF ensembles described before. All NNet modeling was performed using the Neural Network Toolbox in MATLAB (Demuth and Beale, 1992).

Evaluation Criteria

First, we optimized two design-parameters of the k-NN technique, such as the number of selected soils (k) and the power term applied in the weighing system (p). In this phase, different k-NN 'models' were compared with each other, but not to any NNet models. We used only one performance measure for such purposes, the RMSR of the estimations, which is calculated as

$$\mathrm{RMSR} = \sqrt{(1/N) \sum_{i=1}^{N} (\theta_i - \hat{\theta}_i)^2} \quad [7]$$

where N is the number of estimated and measured values, θ and $\hat{\theta}$ are measured and estimated water contents, respectively.

Once the optimal setup for the k-NN technique has been found, we used the k-NN model(s) with the optimal settings as well as the calibrated NNet models to estimate water retention at −33 and −1500 kPa matric potentials for the test data sets. Model performance was evaluated using two measures. We used the MRs as well as the above-defined RMSR to compare the estimation accuracy of the different models. The MR can quantify systematic errors between measurements and estimations and the RMSR can give the accuracy of the estimations in terms of standard deviations. The MR is calculated as

$$\mathrm{MR} = (1/N)\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i) \qquad [8]$$

where all variables are the same as defined in Eq. [7] above.

RESULTS

Optimizing the k and p Terms

We first optimized two design-parameters of the k-NN technique that were introduced in the *Materials and Methods* section. One of the parameters was the number of soils, *k*, to be selected from the reference data set that are then used to formulate the estimate of the output attribute of the target soil. The other parameter was the *p* term introduced in Eq. [6], which is used to weigh each of the selected *k* soils while forming the estimate of the output attribute.

We searched for the optimal parameter settings by gradually changing both of the above parameters in the algorithm and making estimations for the test data set using data of the reference data set. We assigned values to parameter *k* from 1 to 50, with increments of 1 and parameter *p* has been varied between 0 and 3, with increments of 0.1. To avoid possible bias toward one or another input attribute set, we used all four levels of input information to estimate both output attributes (θ33 and θ1500) and averaged their RMSR. Figure 1 shows the average RMSR values obtained using the different combinations of *k* and *p* values with the 1600-sample data set. For this application and this data set, the technique does not seem to be very sensitive to the choice of *p*, differences along *p* are relatively insignificant. It is not very sensitive to the choice of *k* either, as long as *k* is above a certain minimum, 8 or 9 in this case. Choosing zero for *p*—meaning that no relationship between distance and assigned weight is assumed—affects the estimation accuracy negatively, but not in a very significant manner; it is equally a disadvantage to over emphasize the influence of the absolute nearest soil(s) (that is, using large *p*). Figure 2 shows more details about the averaged minimum value of RMSR and the combination of *k*'s and *p*'s used to obtain that. The first contour plot (a) shows the RMSR values. The minimum value for RMSR was 0.04133 (in m^3m^{-3}). It

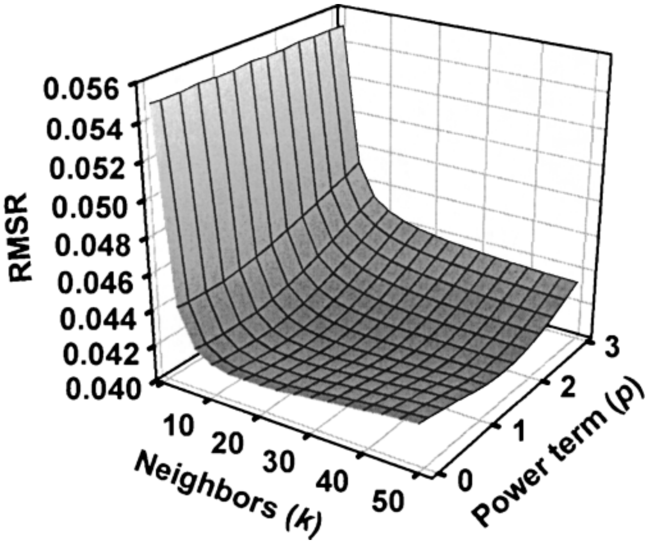

Fig. 1. Three-dimensional representation of the relationship between the number of selected neighbors, the *p* term used in weighing the selected neighbors and the obtained average root-mean-squared residuals, using 1600 soils for each replicate of estimations.

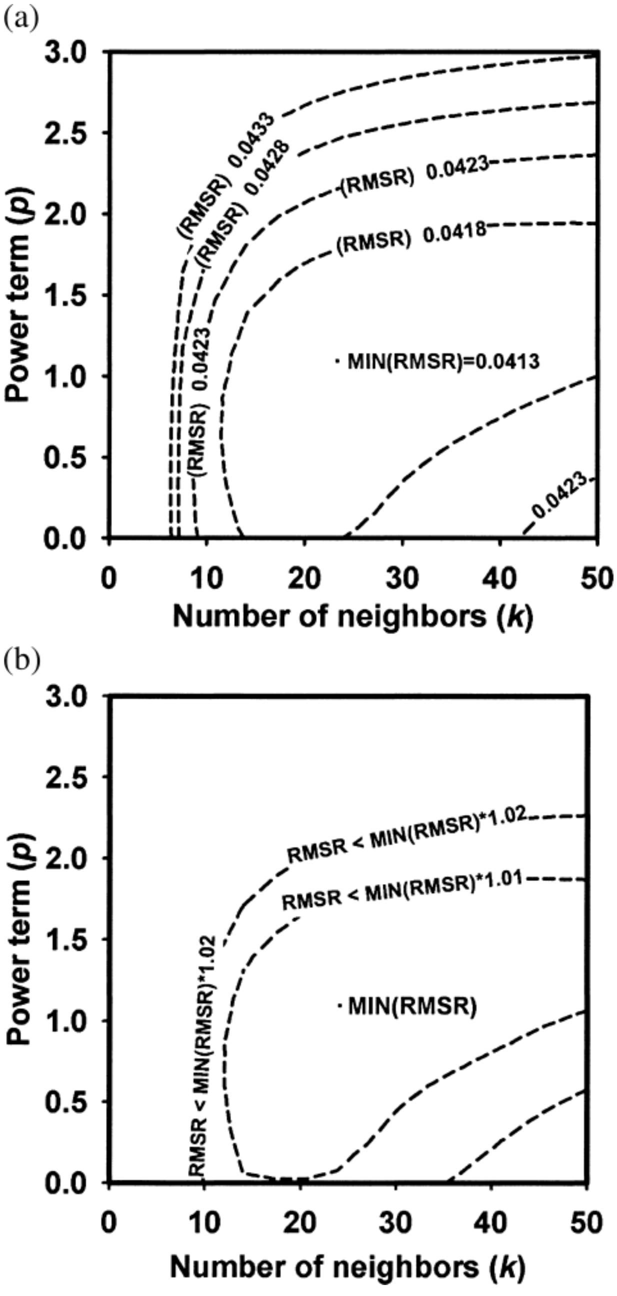

Fig. 2. Two-dimensional representation of the sensitivity of the k-NN technique to suboptimal settings in the number of selected neighbors and the *p* term used in weighing the selected neighbors (a) absolute and (b) relative, using 1600 soils for each replicate of estimations.

can be seen that for a wide range of *k*-*p* combinations the RMSR does not differ much from the minimum value. Plot (b) shows that for values of *k* ranging from 13 to 50, and *p* ranging from 0 to 1.8, most combinations yield RMSRs that are only less than 1% off the minimum RMSR value. The relative insensitivity of a k-NN model to the choice of *k* has also been found before for example, by Lall and Sharma (1996) and Jagtap et al. (2004).

Despite of the insensitivity shown above, one needs to choose a combination of *k* and *p* values to be able to run this algorithm, and the most logical choice is the one leading to minimum RMSR. However, such choices may be affected by the size of the reference data set. For this reason, we also ran the above analysis for the other reference data set sizes: 100, 200, 400, 800. We sorted the combinations of *k* and *p* according to the resulting RMSRs. The *k* and *p* values of the best 10 models were averaged, and are shown for each reference data set size in Fig. 3. For both parameters we found a decreasing trend with decreasing data set size. The optimal value of *k* changed significantly with data set size. The optimal value of *p*, although changed with data set size, remained between 0.95 and 1.10 for the examined data set sizes. The obtained values for *p* were not significantly different from each other. The fact that values of *p* remained around 1 suggests that assuming a simple inverse relationship between the selected samples' weight and distance from the target seems to be a safe first approximation for all data set sizes within the examined range.

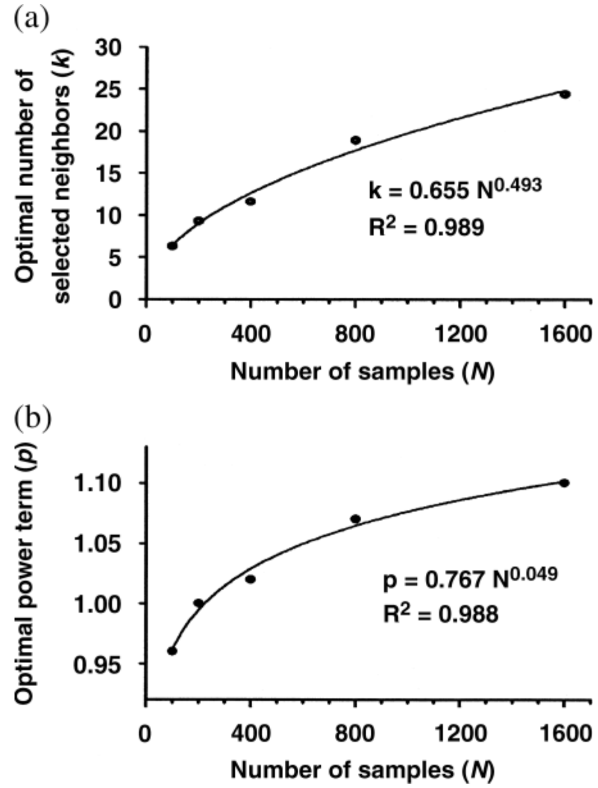

Fig. 3. Effect of data set size on (a) the optimal choice of the number of selected neighbors, and (b) the *p* term used in weighing the selected neighbors.

We found a power function to provide the best fit among the known and simple curve forms to describe the relationship between sample size and each of the two parameters (Fig. 3). The curves fit the data points well. Values for *k* obtained from the fitted curve were approximately $0.62n^{1/2}$ for all data set sizes that we have worked with; approximately two-thirds of the values recommended by Lall and Sharma (1996). For the sake of simplicity and for future reference we included the two approximating power functions shown in Fig. 3 into the algorithm to calculate the optimal settings of *k* and *p*. Subsequent calculations were run using these settings. The optimal value for both parameters were then calculated from the number of soils in the reference data set, and rounded to the nearest integer (*k*) or to the nearest number with two decimals (*p*). We note, however, that the relationships between *n*, *k*, and *p* were set empirically using our data, and may not be optimal when the approach is used with other data sets.

Comparison of k-NN and NNet Models

Using the above settings, we ran the k-NN models, using all five data set sizes, four input attribute sets and made estimations for the two output attributes. We performed the same estimations on identical data also using NNet models. Results, in terms of RMSR, are summarized in Table 2. Root-mean-squared residual values are shown separately for each output attribute, method, input level and each development data set size. Estimations of θ1500 are more accurate than of θ33. One of the potential reasons for this is the larger values of θ33 on average (Table 1), apart from a few outlying high values for θ1500. Another potential reason for this is that θ33 is more influenced by soil structure than θ1500, and soil structure is only represented in the models in an indirect way by D_b. As described before, these RMSR values were calculated by averaging 200 values, resulting from estimations of 200 PTF ensembles each of which used randomly selected data. There is a slight improvement in the estimations when more input attributes are used, but the obtained improvement is not statistically significant at the 0.95 probability level, using any of the two methods. Estimation accuracy does get worse with smaller reference data set size, but such differences are, in many cases, not significant, even between $N = 1600$ and $N = 100$. In most cases, as expected, the standard deviation (SD) of the obtained RMSR became larger with smaller data set size, indicating that estimation accuracy does get more dependent on the samples selected into the actual development/reference data set. This applies for both methods. When the two methods are compared pair wise, results are very close. In most cases, the NNet model resulted in smaller average RMSR. The maximum difference was 0.004 m^3m^{-3}, to the advantage of the NNet model, but there were instances, where the k-NN algorithm performed somewhat better. However, there was no single case, where the differences proved to be statistically significant at the 0.95 probability level.

Table 2. Summary of results, in terms of root-mean-squared residuals (in m^3m^{-3}), for the k-Nearest Neighbor technique with optimized settings and the neural network models. (SSC, sand, silt and clay content; D_b, bulk density; OM, organic matter content).

			Sample size of the PTF development data set (*N*)									
			N = 1600		*N* = 800		*N* = 400		*N* = 200		*N* = 100	
Estimated attribute	**Estimation method**	**Input attribute(s)**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**
Water Retention at −33 kPa	Nearest Neighbor	SSC	0.054	(0.003)	0.054	(0.003)	0.055	(0.003)	0.056	(0.003)	0.058	(0.003)
		SSC + D_b	0.051	(0.002)	0.051	(0.002)	0.052	(0.002)	0.054	(0.003)	0.056	(0.003)
		SSC + OM	0.051	(0.002)	0.051	(0.002)	0.052	(0.003)	0.053	(0.003)	0.055	(0.003)
		SSC + D_b + OM	0.050	(0.002)	0.051	(0.002)	0.052	(0.002)	0.053	(0.003)	0.055	(0.003)
	Neural Network	SSC	0.053	(0.003)	0.053	(0.003)	0.054	(0.003)	0.054	(0.003)	0.055	(0.003)
		SSC + D_b	0.052	(0.003)	0.052	(0.003)	0.052	(0.003)	0.053	(0.003)	0.054	(0.003)
		SSC + OM	0.050	(0.002)	0.051	(0.002)	0.052	(0.003)	0.052	(0.003)	0.054	(0.004)
		SSC + D_b + OM	0.049	(0.002)	0.050	(0.002)	0.051	(0.002)	0.052	(0.003)	0.054	(0.003)
Water Retention at −1500 kPa	Nearest Neighbor	SSC	0.037	(0.002)	0.037	(0.002)	0.038	(0.003)	0.040	(0.003)	0.043	(0.004)
		SSC + D_b	0.035	(0.002)	0.037	(0.002)	0.038	(0.003)	0.040	(0.003)	0.043	(0.004)
		SSC + OM	0.035	(0.002)	0.036	(0.003)	0.038	(0.003)	0.040	(0.003)	0.042	(0.004)
		SSC + D_b + OM	0.035	(0.002)	0.036	(0.002)	0.038	(0.003)	0.040	(0.003)	0.043	(0.004)
	Neural Network	SSC	0.036	(0.002)	0.037	(0.002)	0.037	(0.002)	0.039	(0.003)	0.039	(0.003)
		SSC + D_b	0.035	(0.002)	0.036	(0.002)	0.036	(0.002)	0.037	(0.003)	0.040	(0.005)
		SSC + OM	0.034	(0.002)	0.035	(0.002)	0.035	(0.002)	0.037	(0.003)	0.039	(0.004)
		SSC + D_b + OM	0.034	(0.002)	0.035	(0.002)	0.035	(0.003)	0.036	(0.003)	0.039	(0.005)

We evaluated the performance of the models in terms of bias in the estimates (Table 3). The largest values for estimation bias were obtained using the NNet technique to estimate θ1500 (MR = -0.007 m^3m^{-3}), the maximum bias using the k-NN technique was 0.004 m^3m^{-3}. In practically all cases, zero was within one SD of the obtained bias value, a few exceptions were found for the NNet technique only. Bias did not get significantly larger with smaller data set size or with less input used in the models.

Having no significant differences in RMSR and MR with decreasing data set size and with less input attributes used indicates a large degree of stability of the k-NN technique, and insensitivity to those factors. Having no significant differences in those measures in comparison with the appropriate NNet models indicates a good potential to this technique to be applied to estimate soil hydraulic properties. Our results, should however be validated by the application of this technique to other data sets.

Mean residuals show the extent of overall bias in the estimations. However, such bias may not be equally distributed over the input data range; 'partial' bias may exist along some of the input attributes. We examined the correlations between estimation errors and the input attributes of the models, to reveal any systematic distribution of the estimation errors along any of the input attributes that were used. We only show correlations obtained using 1600 samples in the development/reference data set and SSC, D_b, and OM as input (Table 4). Correlations are shown in terms of R^2 of the linear regression between all data pairs. For the NNet model, R^2 always remained at or below 0.001, meaning that the errors are independent from the greatness of the inputted values. The k-NN technique showed somewhat greater R^2 values, but with one exception, it still remained under 0.008. The R^2 value for the correlation of estimation errors with clay content (estimating θ1500) is greater but is still small ($R^2 = 0.029$). The analysis resulted in somewhat different values for other input attribute sets and data set sizes, but those were comparable in the degree of correlations to the shown example.

Table 3. Summary of results, in terms of mean residuals (in m^3m^{-3}), for the k-Nearest Neighbor technique with optimized settings and the neural network models. (SSC, sand, silt and clay content; D_b, bulk density; OM, organic matter content).

			Sample size of the PTF development data set (*N*)									
			N = 1600		*N* = 800		*N* = 400		*N* = 200		*N* = 100	
Estimated attribute	**Estimation method**	**Input attribute(s)**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**
Water Retention at −33 kPa	Nearest Neighbor	SSC	0.000	(0.003)	0.001	(0.003)	0.001	(0.004)	0.001	(0.005)	0.001	(0.006)
		SSC + D_b	0.001	(0.003)	0.002	(0.003)	0.003	(0.003)	0.003	(0.004)	0.003	(0.006)
		SSC + OM	0.002	(0.002)	0.002	(0.003)	0.003	(0.003)	0.004	(0.004)	0.004	(0.006)
		SSC + D_b + OM	0.002	(0.002)	0.003	(0.003)	0.003	(0.003)	0.004	(0.004)	0.004	(0.005)
	Neural Network	SSC	−0.001	(0.003)	−0.001	(0.003)	0.000	(0.004)	0.000	(0.005)	0.000	(0.006)
		SSC + D_b	0.001	(0.004)	0.002	(0.004)	0.001	(0.004)	0.002	(0.005)	0.000	(0.007)
		SSC + OM	−0.001	(0.003)	0.002	(0.003)	0.003	(0.004)	0.000	(0.005)	0.000	(0.007)
		SSC + D_b + OM	0.000	(0.003)	0.000	(0.003)	0.000	(0.004)	0.001	(0.004)	0.000	(0.006)
Water Retention at −1500 kPa	Nearest Neighbor	SSC	0.001	(0.002)	0.001	(0.002)	0.001	(0.003)	0.001	(0.003)	0.002	(0.004)
		SSC + D_b	0.001	(0.002)	0.001	(0.002)	0.001	(0.003)	0.002	(0.003)	0.002	(0.004)
		SSC + OM	0.001	(0.002)	0.002	(0.002)	0.002	(0.003)	0.003	(0.003)	0.004	(0.004)
		SSC + D_b + OM	0.002	(0.002)	0.002	(0.002)	0.003	(0.003)	0.003	(0.003)	0.004	(0.004)
	Neural Network	SSC	−0.001	(0.002)	−0.005	(0.003)	−0.002	(0.003)	−0.007	(0.005)	−0.001	(0.005)
		SSC + D_b	−0.001	(0.002)	−0.003	(0.003)	−0.002	(0.004)	−0.001	(0.004)	−0.003	(0.005)
		SSC + OM	−0.001	(0.002)	−0.003	(0.003)	−0.001	(0.003)	−0.001	(0.004)	−0.007	(0.006)
		SSC + D_b + OM	0.000	(0.002)	−0.004	(0.004)	−0.002	(0.003)	−0.001	(0.003)	−0.002	(0.005)

Table 4. Correlations (R^2) between estimation errors and the various input attributes of the models that were developed from the data sets with 1600 soils, using soil texture, bulk density and organic matter content as input.

	Nearest Neighbor		Neural Network	
	θ 33	θ 1500	θ 33	θ 1500
Sand	0.00025	0.00139	0.00004	0.00024
Silt	0.00050	0.00543	0.00003	0.00001
Clay	0.00356	0.02953	0.00033	0.00055
Bulk density	0.00734	0.00001	0.00040	0.00072
Organic Matter	0.00229	0.00094	0.00032	0.00031

DISCUSSION

We introduced and tested a k-NN algorithm to serve as a soil hydraulic PTF. Even though the technique is called nonparametric, it takes advantage of a number of design-parameters that need to be optimized for the type of task, before applications. They can be called design-parameters, as they are determined before and independent of applying the algorithm; by application time they are already built in the algorithm. They reflect certain user decisions, rather than represent real parameters. The number of such parameters and the complexity of such optimization tasks may vary greatly depending on the type of problem to be solved. We used a relatively simple algorithm, which we assumed to be able to make estimations efficiently. We made experiments with two of such design-parameters, the number of selected nearest neighbors, k, and the weighing between selected nearest neighbors, in our case represented by the p term. The performance of the algorithm did not depend greatly on k and p. A wide range of suboptimal values around the optimal values yielded only marginal loss in terms of estimation accuracy. There seems to be no need for the user to readjust such parameters to adapt the technique if local soil data are added. Within reasonable limits, having suboptimal settings for k and p would still not result in a major loss in the accuracy of estimations. These settings should, however, be tested/validated for substantially different data sets.

The optimal settings for design parameters k and p varied with the size of the reference data set. Given by the randomized data selection, the distribution of soils in the smaller data sets was similar to that in the largest data sets. What changed was the density of data in the covered data space. Smaller optimal k while having a smaller number of soils in the reference data set indicates that the algorithm gives preference to pick fewer but locally significant information, rather than a wider range of 'general' information. In the meantime, the decrease in the optimal value of p seems to balance this effect. A smaller value of p means that there is—in relative terms—less weight given to the nearest soils, compared with soils that are more distant, but were still selected among the nearest k soils. In practical terms it means, that from a smaller data set, fewer instances will be selected, but those are balanced more equally. It is the opposite for larger data sets. This indicates that the best choice algorithm settings do not prefer going too local and specific to allow, potentially, only the influence of a single instance on the output value. Results confirmed (Fig. 1) that selecting a single soil as accountable nearest neighbor (i.e., $k = 1$) leads to unreliable estimations, and should therefore be avoided.

To define the influence of each selected soil in deriving the final output of the algorithm, some type of weighing was needed. Lall and Sharma (1996) introduced a 'rank based' weighing system within the selected neighbors (Eq. [4]). Application of their method means that for each query, the nearest pick from the reference data set will always have the same weight, the second nearest will have a smaller, but similarly set value, etc., given that the number of selected neighbors, k, is not changed in the meantime. We introduced a different system that does consider the actual distances, d, of the selected neighbors. Such system allows case-by-case weighing. The difference between our system and that of Lall and Sharma (1996) is not expected to be large in localities where the data space is densely populated, but may be larger in the data space where known instances in the reference data set are scarce. We did not compare the two systems, but believe that the k-NN technique—if no other differences are applied—would be similarly insensitive to such differences in design-parameter settings, as it was to other setting changes that we discussed before. However, in case data density is low in some parts of the data space, the situation may belong to one of the extremes the technique can be exposed to, and we suspect that the performance may significantly differ locally as a result of the two different weighing systems.

The importance of the input attributes may vary across the data space. To give an example, a difference of 1% in OM content may have more significance when the target soil has 0% OM content, than when it has 10% OM content. Such had been recognized for example, by Aha and Goldstone (1992). Researchers have since proposed numerous variations of k-NN in an effort to improve its effectiveness on difficult tasks, many of which involve the use of local attribute weighing systems. Examples for the application of such systems are shown in for example, Atkeson et al. (1997) and Howe and Cardie (1997). Local weighing schemes, where attribute weights can vary from instance to instance or input value to input value may perform better in some applications. We think that such variable weights may depend, in part, on database properties and also on local data density. Therefore, in this study we used attribute weights globally, frozen over the entire data space, and left the issue of variable (local) weights to be the subject of future research.

We also experimented with the size of the development/reference data set. We did not obtain significant differences in RMSR or MR with smaller reference data set sizes, which indicate a large degree of stability and insensitivity of the technique to that matter. When the user is not in possession of a large data set, the loss in estimation performance by using a smaller data set with similar data range does not seem to be significantly larger than the loss with the ANN technique in the same situation. This observation however may again be data set dependent and needs to be tested on alternative data sets.

We examined the correlation between estimation errors and the input attributes of the models, to reveal any systematic distribution of the estimation errors along any of the input attributes. In case significant ‘partial’ bias exists, one should introduce additional adjustments to the estimations to compensate for that. Estimation errors were not significantly biased (systematic) along any of the input attributes. Having small or no correlations between those variables mean that the estimation errors are distributed independently from the particular input attributes; there is no ‘partial’ bias in the estimation procedure. It is indicated that this relatively simple form of the k-NN approach is capable to produce efficient estimations in terms of the estimation errors being randomly distributed along the input attributes.

The k-NN technique appears to be a competitive alternative to other techniques to develop PTFs. The statistics of the estimations resembled those obtained using NNet models and, in statistical terms, results did not significantly differ from each other in practically any ways that we examined. Such small differences are unlikely to have significant impact on simulation results that use one water retention estimate or the other. In an earlier study, Nemes et al. (2003) found simulation results to be only marginally affected by more significant differences in the soil hydraulic input data.

An important advantage of this nonparametric algorithm is that, should new data become available, the user is able to include those in the reference data set without the need to redevelop or republish any equations or calculation matrices developed from the original data set. A user will presumably be able to improve estimations for specific local data by incorporating existing local information in the reference data set, without affecting estimations for other sections of the data space. Such would be the case, for instance, if soils with sandy clay texture would be added to most data collections originating from areas of temperate climate zones. One may include data of a set of locally specific sandy clay soils originating from tropical areas, and still use the same temperate climate data set for estimations. Given by the design of the k-NN technique, addition of sandy clay soils will have impact only locally in the sandy clay (and potentially closely neighboring part of the) data space, without causing alteration or degradation in performance for other textural types. Data density is improved locally in the data space. If estimations are made for new sandy clay instances, the nearest soils will be found among the soils that did not preexist in the original data set, but were added later by the local user. For soils with other texture, the original data set will provide the same estimations as before. If the same situation is encountered with parametric PTFs, the local user either needs to develop his/her own independent PTF—for which there may not be enough data—, or needs to add the data and redevelop/readjust the original PTF. In the latter case, however, addition of data with differing data characteristics will change the final form of the relationships between inputs and the output. Such relationships—that is, the equations of parametric PTFs—are valid globally, for the entire range of soils, on which they were developed. This way, estimations made for the entire range of soils are affected by the addition of some specific (e.g., sandy clay) soils. Because the k-NN technique performs all estimation calculations real-time, new data can be added to, or if desired, some of the old data be deleted from the reference data set at any time. In latter case, estimations in the neighborhood of the data space of the discarded data might be negatively affected, but we did not test that in the present study. It has to be considered, however, that if a data set is changed, the previously optimized k and p values may no longer be optimal. While a small change or update in a data set is not expected to change the optimal k and/or p significantly, substantial change to the data set may cause the optimal k and p values to shift significantly. However it is pointed out in Fig. 2 that, at least for this data set, a large variation in k and p values resulted in only an insignificant loss in RMSR.

Points of potential future research include use and comparison of more advanced weighing systems for the entire model globally, or the application of local weighing schemes for different parts of the data space. We also recommend experimenting with other data sets, primarily the application of the technique to soils that originate from different data distribution than the reference data set; and testing the settings of the k and p parameters for other data sets. Dependence of estimation accuracy and uncertainty on actual distance (d) between the target and its nearest neighbors could also be examined, to have a better understanding of potential limitations to this algorithm in practice. Larger distances would be experienced for soils that are underrepresented in the reference data set.

CONCLUSIONS

A k-NN type algorithm has been introduced to make estimations of water contents at −33- and −1500-kPa matric potentials. The performance of the approach has been compared with NNets that were developed using the same data and inputs. The k-NN technique provided estimation statistics, in terms of RMSR and MR that were not significantly different from those obtained using NNets. Performance of the technique—similarly to NNets—showed little sensitivity to using different input attribute sets, or decreased data set sizes used to make the estimations, as well as to certain design-parameter settings. The k-NN technique provides an efficient tool for estimating missing soil water retention data for use in applications in different fields. Literature lists a number of advantages of using such non-parametric approaches over parametric approaches. Our study suggests that to obtain those advantages the user would not necessarily have to compromise estimation accuracy.

REFERENCES

Aha, D.W., and R.L. Goldstone. 1992. Concept learning and flexible weighting. p. 534–539. *In* Proc. 14th Annual Conf. Cognitive Science Society, Bloomington, IN. The Cognitive Science Society, Lawrence Erlbaum Assoc., Wheatridge, CO.

Atkeson, C.G., A.W. Moore and S. Schaal. 1997. Locally weighed learning. Artif. Intell. Rev. 11:11–73.

Clark, M.P., S. Gangopadhyay, D. Brandon, K. Werner, L. Hay, B. Rajagopalan, and D. Yates. 2004. A resampling procedure for generating conditioned daily weather sequences. Water Resour. Res. 40(2):W04304 doi.10.1029/2003WR002747.

Cornelis, W.M., J. Ronsyn, M. van Meirvenne, and R. Hartmann. 2001. Evaluation of pedotransfer functions for predicting the soil moisture retention curve. Soil Sci. Soc. Am. J. 65:638–648.

Dasarathy, B.V. (ed.) 1991. Nearest neighbor (NN) Norms: NN pattern classification techniques. IEEE Computer Society Press, Los Alamitos, CA.

Demuth, H., and M. Beale. 1992. Neural network toolbox manual. MathWorks Inc. Natick, MA.

Efron, B., and R.J. Tibshirani. 1993. An introduction to the bootstrap. Monographs on statistics and applied probability. Chapman and Hall, New York.

Gijsman, A.J., S.S. Jagtap, and J.W. Jones. 2003. Wading through a swamp of complete confusion: How to choose a method for estimating soil water retention parameters for crop models. Eur. J. Agron. 18:77–106.

Harrold, T.I., A. Sharma, and S.J. Sheather. 2003a. A nonparametric model for stochastic generation of daily rainfall occurrence. Water Resour. Res. 39:1300 doi.10.1029/2003WR002182.

Harrold, T.I., A. Sharma, and S.J. Sheather. 2003b. A nonparametric model for stochastic generation of daily rainfall amounts. Water Resour. Res. 39:1343 10.1029/2003WR002570.

Haykin, S. 1994. Neural Networks, a comprehensive foundation. 1st ed. Macmillan College Publishing Company, New York.

Hecht-Nielsen, R. 1990. Neurocomputing. Addison-Wesley, Reading, MA.

Howe, N., and C. Cardie. 1997. Examining locally varying weights for nearest neighbor algorithms. p. 455–466. *In* D. Leake and E. Plaza (Ed.): Proc. 2nd Intl. Conf. on case-based reasoning, Brown University, Providence, RI. Springer-Verlag, New York.

Imam, B., S. Sorooshian, T. Mayr, M.G. Schaap, J.H.M. Wösten, and R.J. Scholes. 1999. Comparison of pedotransfer functions to compute water holding capacity using the van Genuchten model in inorganic soils—Report to IGBP-DIS Soil Data Tasks. IGBP-DIS Working Paper #22. IGBP-DIS, Toulouse, Cédex, France.

Jagtap, S.S., U. Lall, J.W. Jones, A.J. Gijsman, and J.T. Ritchie. 2004. Dynamic nearest-neighbor method for estimating soil water parameters. Trans. ASAE 47:1437–1444.

Karlsson, M., and S. Yakowitz. 1987. Nearest-Neighbor methods for non-parametric rainfall-runoff forecasting. Water Resour. Res. 23: 1300–1308.

Kern, J.S. 1995. Evaluation of soil water retention models based on basic soil physical properties. Soil Sci. Soc. Am. J. 59:1134–1141.

Koekkoek, E.J.W., and H. Booltink. 1999. Neural network models to predict soil water retention. Eur. J. Soil Sci. 50:489–495.

Kumar, D.N., U. Lall, and M.R. Petersen. 2000. Multisite disaggregation of monthly to daily streamflow. Water Resour. Res. 36:1823–1834 10.1029/2000WR900049.

Lall, U., and A. Sharma. 1996. A nearest-neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32:679–693.

Lall, U., B. Rajagopalan, and D.G. Tarboton. 1996. A nonparametric wet/dry spell model for resampling daily precipitation. Water Resour. Res. 32:2803–2823.

Marshall, L., D. Nott, and A. Sharma. 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resour. Res. 40:W02501 10.1029/2003WR002378.

Minasny, B., and A.B. McBratney. 2002. The Neuro-m method for fitting neural network parametric pedotransfer functions. Soil Sci. Soc. Am. J. 66:352–361.

Minasny, B., A.B. McBratney, and K.L. Bristow. 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. Geoderma 93:225–253.

Nemes, A., J.H.M. Wösten, A. Lilly, and J.H. Oude Voshaar. 1999. Evaluation of different procedures to interpolate the cumulative particle-size distribution to achieve compatibility within a soil database. Geoderma 90:187–202.

Nemes, A., M.G. Schaap, and J.H.M. Wösten. 2003. Functional evaluation of pedotransfer functions derived from different scales of data collection. Soil Sci. Soc. Am. J. 67:1093–1102.

Pachepsky, Ya.A., D. Timlin, and G. Várallyay. 1996. Artificial neural networks to estimate soil water retention from easily measurable data. Soil Sci. Soc. Am. J. 60:727–773.

Rajagopalan, B., U. Lall, and D.G. Tarboton. 1996. Nonhomogenous Markov model for daily precipitation. J. Hydrol. Eng. 1: 33–40.

Rajagopalan, B., U. Lall, D.G. Tarboton, and D.S. Bowles. 1997. Multivariate nonparametric resampling scheme for generation of daily weather variables. Stochastic Hydrol. Hydraul. 11:65–93.

Rawls, W.J., T.J. Gish, and D.L. Brakensiek. 1991. Estimating soil water retention from soil physical properties and characteristics. Adv. Soil Sci. 16:213–234.

Sankarasubramanian, A., and U. Lall. 2003. Flood quantiles in a changing climate: Seasonal forecasts and causal relations. Water Resour. Res. 39(5):1134 10.1029/2002WR001593.

Schaap, M.G., and F.J. Leij. 1998. Database-related accuracy and uncertainty of pedotransfer functions. Soil Sci. 163:765–779.

Schaap, M.G., and F.J. Leij. 2000. Improved prediction of unsaturated hydraulic conductivity with the Mualem-van Genuchten model. Soil Sci. Soc. Am. J. 64:843–851.

Schaap, M.G., F.J. Leij, and M.Th. van Genuchten. 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62:847–855.

Schaap, M.G., F.J. Leij, and M.Th. van Genuchten. 2001. ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. J. Hydrol. (Amsterdam) 251: 163–176.

Sharma, A., and U. Lall. 1999. A nonparametric approach for daily rainfall simulation. Math. Comput. Simulat. 48:361–371.

Sharma, A., D.G. Tarboton, and U. Lall. 1997. Streamflow simulation—A nonparametric approach. Water Resour. Res. 33:291–308.

Sharma, A., and R. O'Neill. 2002. A nonparametric approach for representing interannual dependence in monthly streamflow sequences. Water Resour. Res. 38:5 10.1029/2001WR000953.

Soil Survey Staff. 1997. National characterization data. Soil Survey Laboratory, National Soil Survey Center, and Natural Resources Conservation Service, Lincoln, NE.

Souza Filho, F.A., and U. Lall. 2003. Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm. Water Resour. Res. 39:1307 doi:10.1029/2002WR001373.

Tamari, S., J.H.M. Wösten, and J.C. Ruiz-Suárez. 1996. Testing an artificial neural network for predicting soil hydraulic conductivity. Soil Sci. Soc. Am. J. 60:771–774.

Tarboton, D.G., A. Sharma, and U. Lall. 1998. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. Water Resour. Res. 34:107–119.

Tietje, O., and M. Tapkenhinrichs. 1993. Evaluation of pedotransfer functions. Soil Sci. Soc. Am. J. 57:1088–1095.

USDA. 1951. Soil survey manual. U.S. Dep. Agric. Handb. No. 18. U.S. Gov. Print Office, Washington, DC.

van Genuchten, M.Th. 1980. A closed form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44: 892–898.

Wettschereck, D., D.W. Aha and T. Mohri. 1997. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artif. Intell. Rev. 11(1–5):273–314.

Wösten, J.H.M., Ya.A. Pachepsky, and W.J. Rawls. 2001. Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. (Amsterdam) 251: 123–150.

Yakowitz, S. 1993. Nearest-neighbor estimation for null-recurrent Markov time series. Stoch. Proc. Appl. 48:311–318.

Yakowitz, S., and M. Karlsson. 1987. Nearest-Neighbor methods with application to rainfall/runoff prediction. p. 149–160. *In* J.B. Macneil and G.J. Humphries (ed.) Stochastic Hydrology. D. Reidel, Norwell, MA.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek. 2003. A technique for generating regional climate scenarios using a nearest-neighbor algorithm. Water Resour. Res. 39:1199 doi 10.1029/2002WR001769.